Reason in a Changing World

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
(e-mail: pollock@arizona.edu)

A rational agent (artificial or otherwise) residing in a complex changing environment must gather information perceptually, update that information as the world changes, and combine that information with causal information to reason about the changing world. Using the system of defeasible reasoning that is incorporated into the OSCAR architecture for rational agents (Pollock 1995 and 1995a), a set of reason-schemas will be proposed for enabling an agent to perform some of the requisite reasoning. Along the way, solutions will be proposed for the Frame Problem and the Yale Shooting Problem. The principles and reasoning described below have all been implemented in OSCAR.

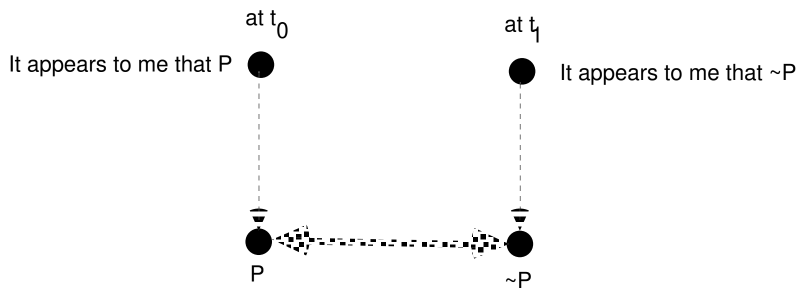
1. Reasoning from Percepts

In a complex changing environment, an agent cannot be equipped from its inception with all the information it needs. It must be capable of gathering new information by sensing its surroundings. This is perception, in a generic sense. Perception is a process that begins with the stimulation of sensors, and ends with beliefs about the agent's immediate surroundings. In perceptual reasoning, the basic inference is from a percept to a conclusion about the world. This enables us to assign propositional contents to percepts. The content of a percept will be taken to be the same as the content of the belief for which the percept provides a reason. But of course, the percept is not the same thing as the belief. Having the percept consists of having a perceptual experience that is *as if* the belief were true. Given this understanding of the content of percepts, we can, as a first approximation, formulate the reasoning from percepts to beliefs as follows (Pollock, 1987):

(1) Having a percept with the content P is a defeasible reason for the agent to believe P.

In this principle, the variable 'P' ranges over the possible contents of percepts. That range will depend upon the perceptual apparatus of the agent in question.

This formulation of the reasoning captures the obvious but important point that perceptual reasoning must be defeasible—appearances can be deceptive. However, that this formulation is not entirely adequate becomes apparent when we consider temporal updating. The world changes, and accordingly percepts produced at different times can support inferences to conflicting conclusions. We can diagram this roughly as follows, where t_1 is a later time than t_0, '$\dashrightarrow$' symbolizes defeasible inference, and ' ' symbolizes defeat relations:



The reasoning seems to produce a case of collective defeat. But it shouldn't. The initial percept supports the belief that P holds *at t_0*, and the second percept supports the belief that P does not hold *at t_1*. These conclusions do not conflict. We can hold both beliefs simply by acknowledging that the world has changed.

We can accommodate this by building temporal reference into the belief produced by perception, and giving the percept a date. This allows us to reformulate the above defeasible reason as follows:

> *PERCEPTION*
> Having a percept at time t with the content P is a defeasible reason for the agent to believe P-at-t.

with the result that the apparent conflict has gone away.

It will be convenient to build the time of the percept into the formula representing the content of the percept. Accordingly, I will adopt the convention of saying that a percept at time t with content P is a percept of P-at-t.

2. Perceptual Reliability

When giving an account of a species of defeasible reasoning, it is as important to characterize the defeaters for the defeasible reasons as it is to state the reasons themselves. This paper assumes the theory of defeasible reasons and reasoning implemented in OSCAR and described in detail in Pollock 1995 and 1995a. One of the central doctrines of that theory is that there are just two kinds of defeaters—*rebutting defeaters* and *undercutting defeaters*. Any reason for denying P-at-t is automatically a rebutting defeater for *PERCEPTION*. An undercutting defeater for an inference from a belief in P to a belief in Q attacks the connection between P and Q rather than merely denying the conclusion. An undercutting defeater is a reason for the formula (P ⊗ Q) (read "It is false that P would not be true unless Q were true", or abbreviated as "P does not guarantee Q"). It is desirable to represent undercutting defeaters for *PERCEPTION* similarly, but we cannot use the content of the percept as the first formula, because that would give us (P ⊗ P), which is logically false. In order to have a formula to represent the content of the undercutting defeater, I will write it as ((It appears to me that P) ⊗ P). The only obvious undercutting defeater is a reliability defeater, which is of a general sort applicable to all defeasible reasons:

> *PERCEPTUAL-RELIABILITY*
> Where r is the strength of *PERCEPTION* , and s < r, "R-at-t, and the probability is less than or equal to s of P's being true given R and that I have a percept with content P" is a defeasible reason for ((It appears to me that P-at-t) ⊗ P-at-t).

3. Temporal Projection

The reason-schema *PERCEPTION* enables an agent to draw conclusions about its current surroundings on the basis of its current percepts. However, that is of little use unless the agent can also draw conclusions about its current surroundings on the basis of earlier (at least fairly recent) percepts. For instance, imagine a robot whose task is to visually check the readings of two meters and then press one of two buttons depending upon which reading is higher. This should not be a hard task, but if we assume that the robot can only look at one meter at a time, it will not be able to acquire the requisite information about the meters using only the reason-schema *PERCEPTION*. The robot can look at one meter and draw a conclusion about its value, but when the robot turns to read the other meter, it no longer has a percept of the first and so is no longer in a position to hold a justified belief about what that meter reads *now*. In order to perform this task, the robot needs some basis for believing that the first meter still reads what it read a moment ago. In other words, the robot must have some basis for regarding the meter reading as a *stable property*—one that tends not to change quickly over time.

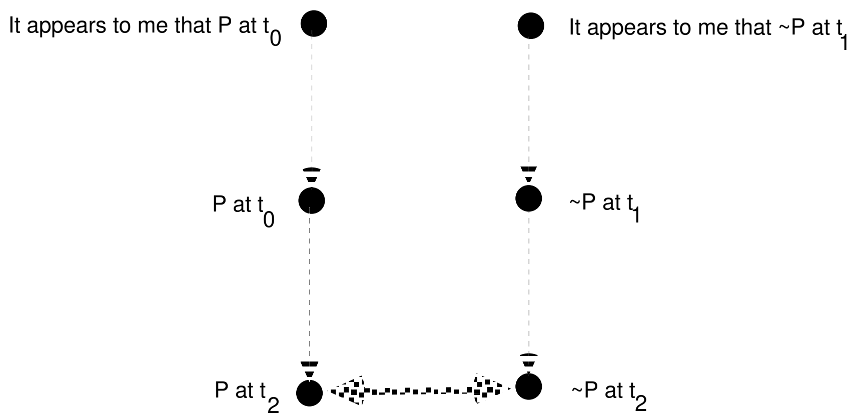
It is natural to suppose that a rational agent, endowed with the ability to learn by induction, can discover inductively that some properties (like the meter readings) are stable to varying degrees, and can bring that knowledge to bear on tasks like the meter-reading task. However, I argued long ago (Pollock 1974) that such inductive learning is not epistemically possible—it presupposes the very stability that is the object of the learning. The argument for this somewhat surprising conclusion is as follows. To say that a property is stable is to say that objects possessing it tend to retain it. To confirm this inductively, an agent would have to re-examine the same object at different times and determine whether the property has changed. The difficulty is that in order to do this, the agent must be able to reidentify the object as the same object at different times. Although this is a complex matter, it seems clear that the agent makes essential use of the perceptible properties of objects in reidentifying them. If all perceptible properties fluctuated wildly, we would be unable to reidentify anything. If objects tended to exchange their perceptible properties abruptly and unpredictably, we would be unable to tell which object was which.[1] The upshot of this is that it is epistemically impossible to investigate the stability of perceptible properties inductively without presupposing that most of them tend to be stable. If we make that general supposition, then we can use induction to refine it by discovering that some perceptible properties are more stable than others, that particular properties tend to be unstable under specifiable circumstances, etc. But our conceptual framework must include a general presumption of stability for perceptible properties before any of this refinement can take place. In other words, the built-in epistemic arsenal of a rational agent must include reason-schemas of the following sort for at least some choices of P:

(2) If $t_0 < t_1$ then believing P-at-t_0 is a defeasible reason for the agent to believe P-at-t_1.

Principle (2) amounts to a presumption that P's being true is a stable property of a time (i.e., a stable fluent, to use the jargon of the situation calculus). A stable property is one such that if it holds at one time, the probability is high that it will continue to hold at a later time. Let ρ be the probability that P will hold at time t+1

[1] A more detailed presentation of this argument can be found in Pollock 1974.

given that it holds at time t. Assuming independence, it follows that the probability that P will hold at time t_1 given that it holds at time t_0 is expt(ρ,t_1–t_0) (where expt(x,y) is x^y). In other words, the strength of the presumption that a stable property will continue to hold over time decays as the time interval increases. This is important for understanding the logic of reasoning about stable properties. To illustrate, consider what I will call *the perceptual updating problem.* Suppose an agent has a percept of P at time t_0, and a percept of ~P at a later time t_1. What an agent *should* conclude (defeasibly) under these circumstances is that the world has changed between t_0 and t_1, and although P was true at t_0, it is no longer true at t_1 and hence no longer true at a later time t_2. If we attempt to reconstruct this reasoning using principle (2), we do not seem to get the right answer. Principle (2) produces the the following inference-graph, and it is a straightforward case of collective defeat. That is, there is one status assignment in which ~P-at-t_2 is undefeated and P-at-t_2 is defeated, and a second status assignment in which ~P-at-t_2 is defeated and P-at-t_2 is undefeated. This is intuitively incorrect.



The solution to getting the reasoning to come out right is to embrace the observation that the strength of the presumption that a stable property will continue to hold over time decays as the time interval increases, and build this into principle (2). The strength of the reason provided by principle (2) must decrease as (t_1 – t_0) increases. This will have the result that the support for ~P-at-t_2 is greater than the support for P-at-t_2, and hence the latter is defeated but the former is not. I propose then that we take the reason-strength of temporal projection to have the form

$$\text{expt}(\rho,(t_1 - t_0)) \cdot \text{degree}(\text{P-at-}t_0)$$

where exp(x) is e^x and degree(P-at-t_0) is the undefeated-degree-of-support of P-at-t_0 and ρ is a constant I will call *the temporal-decay factor.* The reason expt(ρ,(t_1 – t_0)) is multiplied by degree(P-at-t_0) is to ensure that the degree of justification of the conclusion decays uniformly with time. If the reason-strength were greater than degree(P-at-t_0), then the degree of justification of the conclusion would remain constant at degree(P-at-t_0) until the reason-strength decreased to that level. To avoid that, I multiply expt(ρ,(t_1 – t_0)) by degree(P–at-t_0) so that the reason-strength starts out as equal to degree(P–at-t_0) and then decays from there.

The proposal is then that we reformulate (2) as follows:

<u>*TEMPORAL-PROJECTION*</u>

If $t_0 < t_1$, believing P-at-t_0 is a defeasible reason of strength expt(ρ,($t_1 - t_0$)) · degree(P-at-t_0) for the agent to believe P-at-t_1.

This can be illustrated by running OSCAR on the perceptual updating problem. Suppose an object looks red initially, but later looks blue. Lacking any special knowledge about the case, it is reasonable to infer that it was originally red, but later become blue. That is, later perceptual information must take precedence over earlier perceptual information. The key to making the right inferences here is the decaying strength of temporal projections:

```
======================================================================
Problem number 6:  This is the perceptual updating problem.  First, Fred looks red to me.  Later, Fred
looks blue to me.  What should I conclude  about the color of Fred?

Forwards-substantive-reasons:
        *perception*

Backwards-substantive-reasons:
        *temporal-projection*
        *incompatible-colors*

Inputs:
        (the color of Fred is red) : at cycle 1 with justification 1.0
        (the color of Fred is blue) : at cycle 30 with justification 1.0

Ultimate epistemic interests:
    (? x)((the color of Fred is x) at 50)    degree of interest = 0.5
===========================================================================

                                # 1
                                interest: ((the color of Fred is #:^@\y0) at 50)
                                This is of ultimate interest
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is red) at 1)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 1
 It appears to me that ((the color of Fred is red) at 1)
 # 2
 ((the color of Fred is red) at 1)
 Inferred by:
            support-link #1 from { 1 } by *perception*
 undefeated-degree-of-support = 0.9
 # 3
 ((the color of Fred is red) at 50)
 undefeated-degree-of-support = 0.8569420073847128
 Inferred by:
            support-link #2 from { 2 } by *temporal-projection*  defeaters: { 7 }
 This discharges interest 1
                                # 5
                                interest: ~((the color of Fred is red) at 50)
                                Of interest as a defeater for support-link 2 for node 3
            =========================================
            Justified belief in ((the color of Fred is red) at 50)
            answers #<Query 1: (? x)((the color of Fred is x) at 50)>
            =========================================
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is blue) at 30)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 4
 It appears to me that ((the color of Fred is blue) at 30)
 # 5
 ((the color of Fred is blue) at 30)
 Inferred by:
            support-link #3 from { 4 } by *perception*
 undefeated-degree-of-support = 0.9
 # 6
 ((the color of Fred is blue) at 50)
 Inferred by:
            support-link #4 from { 5 } by *temporal-projection*  defeaters: { 8 }
 undefeated-degree-of-support = 0.8821699783465813
 This discharges interest 1
                                # 9
```

```
                        interest: ~((the color of Fred is blue) at 50)
                        Of interest as a defeater for support-link 4 for node 6
          ========================================
          Justified belief in ((the color of Fred is blue) at 50)
          answers #<Query 1: (? x)((the color of Fred is x) at 50)>
          ========================================
 # 7
 ~((the color of Fred is red) at 50)
 Inferred by:
           support-link #5 from { 6 } by *incompatible-colors*
 undefeated-degree-of-support = 0.8821699783465813
 defeatees: { link 2 for node 3 }
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 3> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          ========================================
          Lowering the undefeated-degree-of-support of ((the color of Fred is red) at 50)
          retracts the previous answer to #<Query 1: (? x)((the color of Fred is x) at 50)>
          ========================================
================== ULTIMATE EPISTEMIC INTERESTS ===================
 Interest in (? x)((the color of Fred is x) at 50)
 is answered by node 6:  ((the color of Fred is blue) at 50)
---------------------------------------------------
```

4. Temporal Indexicals

An agent that did all of its temporal reasoning using the reason-schemas described above would be led into crippling computational complexities. Every time it wanted to reuse a belief about its surroundings, it would have to reinfer it for the present time. Inference takes time, so by the time it had reinferred the belief, other beliefs with which the agent might want to combine this belief in further inference would themselves no longer be current. To get around this difficulty, the agent would have to make inferences about some time in the near future rather than the present, inferring a number of properties of that time, and then combine those properties to make a further inference about that time, and finally project that new property into the future for use in further inference. This would not be computationally impossible, but it would make life difficult for an agent that had to reason in this way.

Human beings achieve the same result in a more efficient way by employing the temporal indexical "now". Rather than repeatedly reinferring a property as time advances, they infer it once as holding *now*, and that single belief is retained until it becomes defeated. The mental representation (i.e., formula) believed remains unchanged as time advances, but the content of the belief changes continuously in the sense that at each instant it is a belief about *that instant*. This has the effect of continuously updating the agent's beliefs in accordance with the defeasible reasons of section three, but no actual reasoning need occur.

The use of "now" can be implicit or explicit. That is, we can either write "x is red" or "x is red now". The latter is used primarily for emphasis. It keeps the representation simpler if we drop the temporal reference rather than putting in the "now", so that is the course that will be followed below.

Percepts are always percepts of the agent's present situation, so we can regard them as providing defeasible reasons for inferences about the present:

> *INDEXICAL-PERCEPTION*
> Having a percept at time t with the content P is a defeasible reason for the agent to believe P.

INDEXICAL-PERCEPTION is defeated by considerations of reliability just as *PERCEPTION* is:

> <u>*INDEXICAL-PERCEPTUAL-RELIABILITY*</u>
> Where r is the strength of *INDEXICAL-PERCEPTION*, "R-at-t, and the probability is less than r of P's being true given R and that I have a percept with content P at t" is a defeasible undercutting defeater for *INDEXICAL-PERCEPTION.*

The conclusion P is automatically projected into the future just by retaining it, so *INDEXICAL-PERCEPTION* can be viewed as combining *PERCEPTION* and *TEMPORAL-PROJECTION* into a single reason-scheme.

A complication arises for the reasoner in dealing with conclusions containing an implicit or explicit "now" representing an implicit use of *TEMPORAL-PROJECTION*. The degree-of-support for a conclusion drawn by *TEMPORAL-PROJECTION* decreases as the time interval increases. Retaining a conclusion containing "now" is equivalent to making an inference by *TEMPORAL-PROJECTION*. Accordingly, the degree-of-support for that conclusion must decay over time, just as if it were being continually re-inferred by *TEMPORAL-PROJECTION.* To handle this, we must make a distinction between *temporal* and *atemporal* conclusions, where the former are those containing an explicit or implicit "now". Atemporal conclusions have fixed degrees-of-support, but the degree of support of a temporal conclusion must decay proportionally to expt(ρ,Δt) where Δt is the length of time since it was initially inferred.

These principles can be illustrated by the following problem. First, Fred looks red to me. Later, I am informed by Merrill that I am then wearing blue-tinted glasses. Later still, Fred looks blue to me. All along, I know that the probability is not high of Fred being blue given that Fred looks blue to me, but I am wearing blue-tinted glasses. What should I conclude about the color of Fred? By *INDEXICAL-PERCEPTION* I first infer that Fred is red. Later I infer that Fred is blue, and in accordance with the previously noted solution to the perceptual updating problem, the earlier conclusion that Fred is red is defeated. By I also infer that I am wearing blue-tinted glasses, which by *INDEXICAL-PERCEPTUAL-RELIABILITY* defeats the conclusion that Fred is blue and reinstates the conclusion that Fred is red. OSCAR performs this reasoning as follows:

```
======================================================================
Problem number 8:  First, Fred looks red to me.  Later, I am informed by Merrill that I am then wearing
blue-tinted glasses.  Later still, Fred looks blue to me.  All along, I know that the probability is not high of
Fred being blue given that Fred looks blue to me, but I am wearing blue tinted glasses.  What should I
conclude about the color of Fred?

Forwards-substantive-reasons:
        *indexical-perception*
        *perception*
        *reliable-informant*

Backwards-substantive-reasons:
        *indexical-perceptual-reliability*
        *perceptual-reliability*
        *temporal-projection*
        *indexical-incompatible-colors*

Inputs:
        (the color of Fred is red) : at cycle 1 with justification 0.8
        (Merrill reports that I am wearing blue tinted glasses) : at cycle 20 with justification 1.0
        (the color of Fred is blue) : at cycle 30 with justification 0.8

Given:
        ((the probability of (the color of Fred is blue) given ((I have a percept with content (the color of
Fred is blue)) & I am wearing blue tinted glasses)) <= 0.8)  :  with justification = 1.0
        (Merrill is a reliable informant)  :  with justification = 1.0

Ultimate epistemic interests:
    (? x)(the color of Fred is x)    degree of interest = 0.7
```

```
================================================================================

 # 1
 ((the probability of (the color of Fred is blue) given ((I have a percept with content (the color of Fred is
blue)) & I am wearing blue tinted glasses)) <= 0.8)
 given
 undefeated-degree-of-support = 1.0 at cycle 1.
 # 2
 (Merrill is a reliable informant)
 given
 undefeated-degree-of-support = 1.0 at cycle 1.
                                        # 1
                                        interest: (the color of Fred is #:^@\y0)
                                        This is of ultimate interest
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is red) at 1)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 3
 It appears to me that ((the color of Fred is red) at 1)
 # 5
 (the color of Fred is red)
 Inferred by:
              support-link #4 from { 3 } by *indexical-perception*  defeaters: { 13 }
 undefeated-degree-of-support = 0.7992003998667001 at cycle 2.
 This discharges interest 1
                                        # 5
                                        interest: ~(the color of Fred is red)
                                        Of interest as a defeater for support-link 4 for node 5
              ===========================================
              Justified belief in (the color of Fred is red)
              answers #<Query 1: (? x)(the color of Fred is x)>
              ===========================================
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((Merrill reports that I am wearing blue tinted glasses) at 20)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 6
 It appears to me that ((Merrill reports that I am wearing blue tinted glasses) at 20)
 # 7
 ((Merrill reports that I am wearing blue tinted glasses) at 20)
 Inferred by:
              support-link #5 from { 6 } by *perception*
 undefeated-degree-of-support = 0.9 at cycle 20.
 # 9
 (I am wearing blue tinted glasses at 20)
 Inferred by:
              support-link #7 from { 2 , 7 } by *reliable-informant*
 undefeated-degree-of-support = 0.9 at cycle 22.
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
It appears to me that ((the color of Fred is blue) at 30)
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 # 10
 It appears to me that ((the color of Fred is blue) at 30)
                                        # 13
                                        interest: (I am wearing blue tinted glasses at 30)
                                        For interest 15 by *indexical-perceptual-reliability*
                                        This interest is discharged by node 15
 # 12
 (the color of Fred is blue)                  DEFEATED
 Inferred by:
              support-link #9 from { 10 } by *indexical-perception*  defeaters: { 16 , 14 }
 undefeated-degree-of-support = 0.8 at cycle 30.
 This discharges interest 1
                                        # 15
                                        interest: (((it appears to me that (the color of Fred is blue)) at 30) ⊗ (the color
                                         of Fred is blue))
                                        Of interest as a defeater for support-link 9 for node 12
                                        # 16
                                        interest: ~(the color of Fred is blue)
                                        Of interest as a defeater for support-link 9 for node 12
              ===========================================
              Justified belief in (the color of Fred is blue)
              answers #<Query 1: (? x)(the color of Fred is x)>
              ===========================================
```

```
# 13
~(the color of Fred is red)                DEFEATED
Inferred by:
            support-link #10 from { 12 } by *indexical-incompatible-colors*
undefeated-degree-of-support = 0.7992003998667001 at cycle 31.
defeatees: { link 4 for node 5 }
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           #<Node 5> has become defeated.
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           ==========================================
           Lowering the undefeated-degree-of-support of (the color of Fred is red)
           retracts the previous answer to #<Query 1: (? x)(the color of Fred is x)>
           ==========================================
# 14
~(the color of Fred is blue)
Inferred by:
            support-link #11 from { 5 } by inversion from contradictory nodes 13 and 5
defeatees: { link 9 for node 12 }
undefeated-degree-of-support = 0.0 at cycle 32.
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           #<Node 14> has become defeated.
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           The undefeated-degree-of-support of #<Node 13> has decreased to 0.7984015989338665
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
# 15
(I am wearing blue tinted glasses at 30)
Inferred by:
            support-link #12 from { 9 } by *temporal-projection*
undefeated-degree-of-support = 0.8910448503742513 at cycle 34.
This discharges interest 13
# 16
(((it appears to me that (the color of Fred is blue)) at 30) ⊗ (the color of Fred is blue))
Inferred by:
            support-link #13 from { 1 , 15 } by *indexical-perceptual-reliability*
defeatees: { link 9 for node 12 }
undefeated-degree-of-support = 0.8910448503742513 at cycle 34.
This node is inferred by discharging interest #15
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           The undefeated-degree-of-support of #<Node 14> has increased to 0.7740308476712257
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           The undefeated-degree-of-support of #<Node 5> has increased to 0.7740308476712257
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           #<Node 12> has become defeated.
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           #<Node 13> has become defeated.
           vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
           ==========================================
           Justified belief in (the color of Fred is red)
           answers #<Query 1: (? x)(the color of Fred is x)>
           ==========================================
           ==========================================
           Lowering the undefeated-degree-of-support of (the color of Fred is blue)
           retracts the previous answer to #<Query 1: (? x)(the color of Fred is x)>
           ==========================================
================== ULTIMATE EPISTEMIC INTERESTS ===================
 Interest in (? x)(the color of Fred is x)
 is answered by node 5:  (the color of Fred is red)
---------------------------------------------------
```

This reasoning diagramed in figure 1.

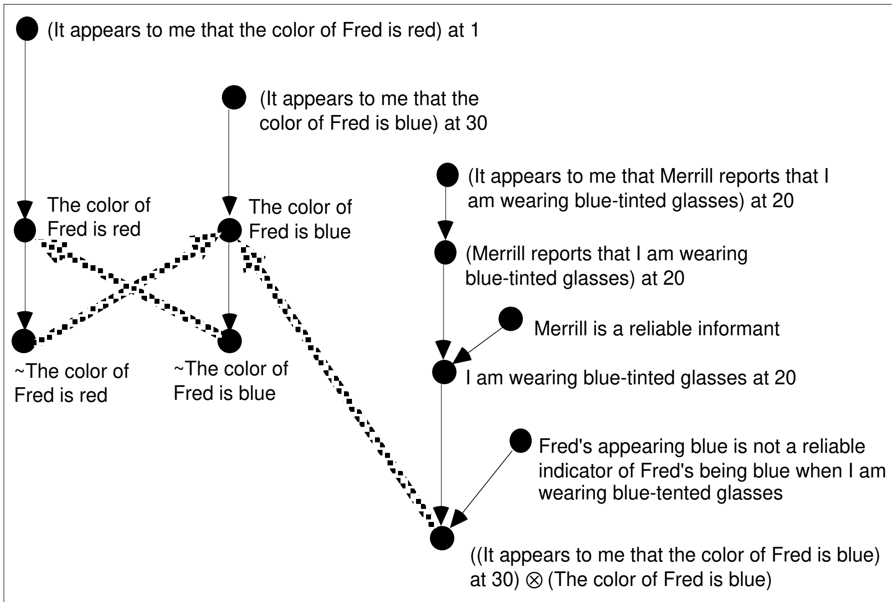


Figure 1. Inference graph

5. The Frame Problem

Reasoning about what will change if an action is performed or some other change occurs often presupposes knowing what will not change. Early attempts to model such reasoning deductively proceeded by adopting a large number of "frame axioms", which were axioms to the effect that if something occurs then something else will not change. For instance, in a blocks world one of the frame axioms might be "If a block is moved, its color will not change". It soon became apparent that complicated situations required more frame axioms than axioms about change, and most of the system resources were being occupied by proofs that various properties did not change. In a realistically complicated situation, this became unmanageable. What became known as the *Frame Problem* is the problem of reorganizing reasoning about change so that reasoning about non-changes can be done efficiently (McCarthy and Hayes, 1969; Janlert, 1987).

AI hackers, as Hayes [1987] calls them, avoided this problem by adopting the "sleeping dog strategy" (Haugeland 1987). Starting with STRIPS, actual planning systems maintained databases of what was true in a situation, and with each possible action they stored lists of what changes those actions would produce. For planning systems intended to operate only in narrowly circumscribed situations, this approach is effective, although for general-purpose planning it quickly becomes unwieldy. In the attempt to provide a more general theory that justifies this approach as a special case, several authors (Sandewall 1972, McCarthy) proposed reasoning about change defeasibly and adopting some sort of defeasible inference scheme to the effect that it is reasonable to believe that something doesn't change unless you are forced to conclude otherwise. But to make the idea work, one needs both a precise framework for defeasible reasoning and a precise formulation of the requisite defeasible inference schemes. That proved to be a difficult problem.

The temporal projection principles defended above can be regarded as a precise formulation of the defeasible inference schemes sought. For the reasons given in the preceding discussion of temporal indexicals, I do not think that these principles by themselves constitute a solution to the Frame Problem. However, I propose that when they are coupled with the additional principles regarding temporal indexicals, an adequate solution to the Frame Problem is forthcoming. These principles have all been implemented in OSCAR.

6. Reasoning about Change

The frame problem concerns efficient reasoning about non-change. This only arises because we must often know what does not change in order to reason about what changes. Steve Hanks and Drew McDermott [1986] were the first to observe that even with defeasible principles of non-change, a reasoner will often be unable to determine what changes and what does not. (This could be viewed as an argument that the frame problem has not yet been solved, but it more convenient to regard this as a separate problem.) They illustrated this with what has become known as "the Yale shooting problem". The general form of the problem is this. Suppose we have a causal law to the effect that if P is true at a time t and action A is performed at that time, then Q will be true shortly thereafter. (More generally, A could be anything that becomes true at a certain time. What is significant about actions is that they are changes.) Suppose we know that P is true now, and Q false. What should we conclude about the results of performing action A in the immediate future? Hanks and McDermott illustrate this by taking P to be "The gun is loaded, in working condition, and pointed at Jones", Q to be "Jones is dead", and A to be the action of pulling the trigger. We suppose (simplistically) that there is a causal law dictating that if the trigger is pulled on a loaded gun that is in working condition and pointed at someone, that person will shortly be dead. Under these circumstances, it seems clear that we should conclude that Jones will be dead shortly after the trigger is pulled.

The difficulty is that all we can deduce from what we are given is that when A is performed either P will no longer be true or Q will be true shortly thereafter. Intuitively, we want to conclude (at least defeasibly) that P will remain true at the time A is performed and Q will therefore become true shortly thereafter. But none of our current machinery enables us to distinguish between P and Q. Because P is now true and Q is now false, we have a defeasible reason for believing that P will still be true when A is performed, and we have a defeasible reason for believing that Q will still be false shortly thereafter. We know that one of these defeasible conclusions will be false, but we have no basis for choosing between them, so this becomes a case of collective defeat. That, however, is the intuitively wrong answer.

When we reason about causal mechanisms, we think of the world as "unfolding" temporally, and changes only occur when they are forced to occur by what has already happened. In our example, when A is performed, nothing has yet happened to force a change in P, so we conclude defeasibly that P remains true. But given the truth of P, we can then deduce that at a slightly later time, Q will become true. Thus when causal mechanisms force there to be a change, we conclude defeasibly that the change occurs in the later states rather than the earlier states. This seems to be part of what we mean by describing something as a causal mechanism. Causal mechanisms are systems that force changes, where "force" is to be understood in terms of temporal unfolding.[2]

When reasoning about such a causal system, part of the force of describing it as causal must be that the defeasible presumption against the effect occurring is somehow removed. Thus, although we normally expect Jones to remain alive, we do not expect this any longer when he is shot. To remove a defeasible presumption is to defeat it. This suggests that there is some kind of general "causal" defeater for the temporal projection principles adumbrated above. The problem is to state this defeater precisely.

As a first approximation we might try:

(4) For every $\varepsilon \geq 0$ and $\delta > 0$, "A&P-at-(t+ε) & (A&P causes Q)" is an undercutting defeater for the defeasible inference from ~Q-at-t to ~Q-at-(t+ε+δ).

This account of causal reasoning requires causation to be temporally asymmetric. That is, "A&P causes Q" means, in part, that if A&P becomes true then Q will *shortly* become true. This precludes *simultaneous* causation, in which Q is caused to be true at t by A&P being true at t. This may seem problematic, on the grounds that simultaneous causation occurs throughout the real world. For instance, colliding billiard balls in classical physics might seem to illustrate simultaneous causation. However, this is a mistake. If two billiard balls collide at time t with velocity vectors pointing towards each other, they do not also have velocity vectors pointing away from each other at the very same time. Instead, this illustrates what I have elsewhere [1984] called *instantaneous* causation. Instantaneous causation requires that if A&P becomes true at t, then for some $\delta > 0$, for every ε such that $0 < \varepsilon \leq \delta$, Q will be true at t+ε.[3] I believe that instantaneous causation is all that is required for describing the real world.

I have followed AI-convention here in talking about causal change in terms of causation. However, that introduces unnecessary complexities. For example, it is generally assumed in the philosophical literature on causation, that if P causes Q then Q would not have been true if P were not true. This has the consequence that in cases in which there are two independent factors each of which would be sufficient by itself to cause the same effect, if both occur then neither cause it. These are cases of causal overdetermination. A familiar example of causal overdetermination occurs when two assailants shoot a common victim at the same time. Either shot would be fatal. The result is that neither shot is such that if it had not occurred then the victim would not have died, and hence, it is generally maintained, neither shot caused the death of the victim. However, this kind of failure of causation ought to be irrelevant to the kind of causal reasoning under discussion in connection with change. Principle (4) ought to apply to cases of causal overdetermination as well as to genuine cases of causation.

I take it (and have argued in my [1984]) that all varieties of causation (including causal overdetermination) arise from the instantiation of "causal laws". These are what I have dubbed *nomic generalizations*, and have discussed at length in my [1990]. Nomic generalizations are symbolized as "P ≡> Q", where P and Q are formulas and '≡>' is a variable-binding operator, binding all free occurrences of variables in P and Q. An informal gloss on "P ≡> Q" is "Any physically-possible P would be a Q". For

[2] This intuition is reminiscent of Shoham's (1987) "logic of chronological ignorance", although unlike Shoham, I propose to capture the intuition without modifying the structure of the system of defeasible reasoning.

[3] I assume that time has the structure of the reals, although that assumption is not required for the implementation.

example, the law that electrons are negatively charged could be written "(x is an electron) ≡> (x is negatively charged)". The free occurrences of 'x' are bound by '≡>'. I propose that we replace "(A&P causes Q)" in (4) by "(A&P ≡> Q will shortly be true)", where the latter typically results from instantiating more general laws. More precisely, let us define "A when P is causally sufficient for Q after an interval ε" to mean

$$(\forall t)[(\text{A-at-t} \ \& \ \text{P-at-t}) \equiv\!> (\exists\gamma)(\forall\delta)(\varepsilon < \delta \leq \gamma \supset \text{Q-at-}(t+\delta))].$$

Instantaneous causation is causal sufficiency with an interval 0.

My proposal is to replace 'causes' by 'causal sufficiency' in (4). Modifying it to take account of the interval over which the causation occurs:

<u>*CAUSAL-UNDERCUTTER*</u>

Where $t_0 \leq t_1$ and $(t_1+\varepsilon) < t$, "A-at-t_1 & Q-at-t_1 & (A when Q is causally sufficient for ~P after an interval ε)" is a defeasible undercutting defeater for the inference from P-at-t_0 to P-at-t by *TEMPORAL-PROJECTION*.

For causal reasoning, we also want to use the causal connection to support inferences to conditionals. This is more complicated than it might initially seem. The difficulty is that, for example,

> (the gun is fired when the gun is loaded) is causally sufficient for ~(Jones is alive) after an interval 20

does not imply that if the gun is fired at t and the gun is loaded at t then Jones is dead at t+20. Recall the discussion of instantaneous causation. All that is implied is that Jones is dead over some interval open on the left and with t+20 as the lower bound. We can conclude that *there is* at time > t+20 at which Jones is dead, but it does not follow as a matter of logic that Jones is dead at any particular time because, at least as far as this causal law is concerned, Jones could become alive again after becoming dead. To infer that Jones is dead at a particular time after t+20, we must combine the causal sufficiency with temporal projection. This yields the following principle:

<u>*CAUSAL-IMPLICATION*</u>

"(A when P is causally sufficient for Q after an interval ε) & A-at-t & P-at-t & $((t+\varepsilon) < t^*)$" is a defeasible reason for "Q-at-t*".

With the use of these principles, it is concluded that Jones is not alive at 50:

```
======================================================================
Problem number 13: This is the Yale Shooting Problem. I know that the gun being fired while loaded
will cause Jones to become dead. I know that the gun is initially loaded, and Jones is initially
alive. Later, the gun is fired. Should I conclude that Jones becomes dead?

Forwards-substantive-reasons:

Backwards-substantive-reasons:
        *indexical-temporal-projection*
        *temporal-projection*
        *indexical-causal-undercutter*
        *indexical-causal-implication*

Start reasoning at cycle 50

Inputs:
```

```
Given:
        ((Jones is alive) at 20)  :  with justification = 1.0
        (the gun is loaded at 20)  :  with justification = 1.0
        (the gun is fired at 30)  :  with justification = 1.0
        (the gun is fired when the gun is loaded is causally sufficient for ~(Jones is alive) after an interval
           10)  :  with justification = 1.0

Ultimate epistemic interests:
    (? (Jones is alive))    degree of interest = 0.75
===============================================================================
Nodes marked DEFEATED have that status at the end of the reasoning.

 # 1
 ((Jones is alive) at 20)
 given
 # 2
 (the gun is loaded at 20)
 given
 # 3
 (the gun is fired at 30)
 given
 # 4
 (the gun is fired when the gun is loaded is causally sufficient for ~(Jones is alive) after an interval 10)
 given
                              # 1
                              interest: (Jones is alive)
                              This is of ultimate interest
                              Of interest as a defeater for support-link 7 for node 7
                              # 2
                              interest: ~(Jones is alive)
                              This is of ultimate interest
                              Of interest as a defeater for support-link 5 for node 5
 # 5
 (Jones is alive)                DEFEATED
 Inferred by:
            support-link #5 from { 1 } by *indexical-temporal-projection*  defeaters: { 8 , 7 }   DEFEATED
 defeatees: { link 7 for node 7 }
 This discharges interest 1
                              # 3
                              interest: (((Jones is alive) at 20) ⊗ (Jones is alive))
                              Of interest as a defeater for support-link 5 for node 5
          =========================================
          Justified belief in (Jones is alive)
          answers #<Query 1: (? (Jones is alive))>
          =========================================
                              # 4
                              interest: (the gun is loaded at 30)
                              For interest 2 by *indexical-causal-implication*
                              For interest 3 by *indexical-causal-undercutter*
                              This interest is discharged by node 6
 # 6
 (the gun is loaded at 30)
 Inferred by:
            support-link #6 from { 2 } by *temporal-projection*
 This discharges interest 4
 # 7
 ~(Jones is alive)
 Inferred by:
            support-link #7 from { 4 , 3 , 6 } by *indexical-causal-implication*  defeaters: { 5 }
 defeatees: { link 5 for node 5 }
 This node is inferred by discharging interest #2
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 7> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          #<Node 5> has become defeated.
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          =========================================
          Lowering the undefeated-degree-of-support of (Jones is alive)
          retracts the previous answer to #<Query 1: (? (Jones is alive))>
          =========================================
 # 8
 (((Jones is alive) at 20) ⊗ (Jones is alive))
 Inferred by:
            support-link #8 from { 4 , 3 , 6 } by *indexical-causal-undercutter*
 defeatees: { link 5 for node 5 }
```

```
 This node is inferred by discharging interest #3
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          The undefeated-degree-of-support of #<Node 7> has increased to 0.9646229185299475
          vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
          ==========================================
          Justified belief in ~(Jones is alive)
          answers #<Query 1: (? (Jones is alive))>
          ==========================================
============================================================


================== ULTIMATE EPISTEMIC INTERESTS ===================
 Interest in (? (Jones is alive))
 is answered by node 7:  ~(Jones is alive)
---------------------------------------------------
```

This reasoning is diagramed in figure 2.

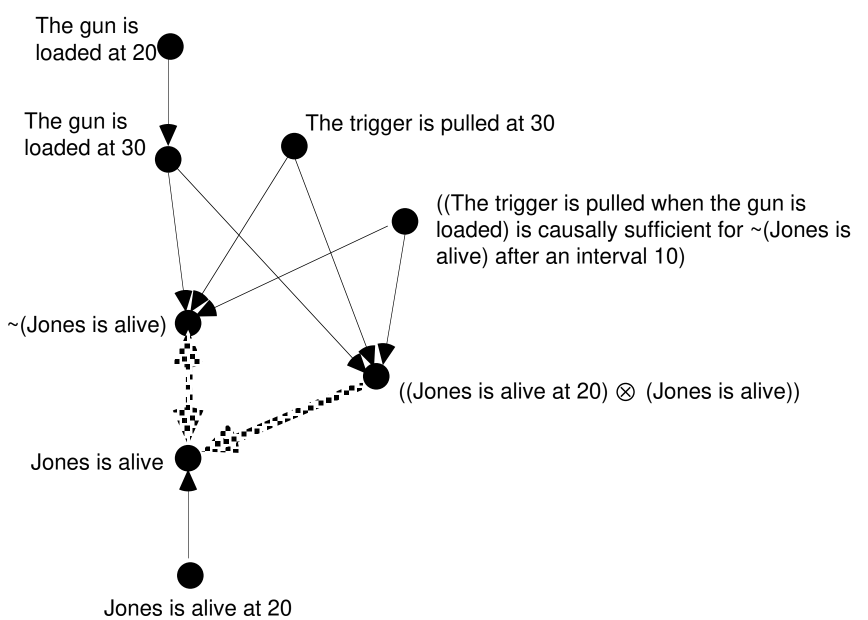


Figure 2. The Yale Shooting Problem

This, as far as I know, is the first *implemented* fully general solution to the Yale Shooting Problem.[4] It is noteworthy how simple it is to implement such principles in OSCAR, making OSCAR a potent tool for epistemological analysis.

[4] Shanahan [1996] describes an implemented solution to special cases of the Yale Shooting Problem that occur in robot navigation. However, his proposed solution is based upon circumscription, and so is in principle not implementable in any but special cases. Kartha and Lifschitz [1995] also describe implemented solutions to special cases, based upon circumscription. It is noteworthy that their implementation works by attempting to reduce the circumscription to a set of first-order frame axioms, which is just what the appeal to nonmonotonic logic and defeasible reasoning was originally intended to avoid.

References

Hanks, Steve, and McDermott, Drew
1986 "Default reasoning, nonmonotonic logics, and the frame problem", AAAI-86.
Haugland, John
1987 "An overview of the frame problem", in Z. Pylyshyn (ed.) *The Robot's Dilemma*,, MIT Press.
Hayes, Patrick
1987 "What the frame problem is and isn't", in Z. Pylyshyn (ed.) *The Robot's Dilemma*,, MIT Press.
Janlert, Lars-Erik
1987 "Modeling change—the frame problem", in Z. Pylyshyn (ed.) *The Robot's Dilemma*,, MIT Press.
McCarthy, John, and Hayes, Patrick
1969 "Some philosophical problems from the standpoint of artificial intelligence". In B. Metzer & D. Michie (eds.), *Machine Intelligence 4*. Edinburgh: Edinburgh University Press.
Kartha, G. Neelakantan, and Lifschitz, Vladimir
1995 "A simple formalization of actions using circumscription". *Proceedings of IJCAI 1995*, 1970-1975.
Pollock, John
1974 *Knowledge and Justification*, Princeton University Press.
1984 *The Foundations of Philosophical Semantics*, Princeton University Press.
1987 *Contemporary Theories of Knowledge*, Rowman and Littlefield.
1990 *Nomic Probability and the Foundations of Induction*, Oxford University Press.
1995 *Cognitive Carpentry*, MIT Press.
1995a *The OSCAR Manual*, available online from http://info-center.ccit.arizona.edu/~oscar.
Sandewall, Erik
1972 "An approach to the frame problem and its implementation". In B. Metzer & D. Michie (eds.), *Machine Intelligence 7*. Edinburgh: Edinburgh University Press.
Shanahan, Murray
1996 "Robotics and the common sense informatic situation", *Proceedings of the 12th European Conference on Artificial Intelligence*, John Wiley & Sons.
Shoham, Yoav
1987 *Reasoning about Change*, MIT Press.